Does Super-Resolution Improve OCR Performance in the Real World ? A Case Study on Images of Receipts

Vivien Robert, Hugues Talbot

DOES SUPER-RESOLUTION IMPROVE OCR PERFORMANCE IN THE REAL WORLD? A CASE STUDY ON IMAGES OF RECEIPTS

Vivien Robert[1] *& Hugues Talbot*[2]

[1]Artefact, 19 rue Richer 75009 Paris, France
[2]Université Paris-Saclay, CentraleSupelec, Inria, 9 rue Joliot-Curie, 91190 Gif sur Yvette, France

ABSTRACT

Recently, many deep learning methods have been used to handle single image super-resolution (SISR) tasks and often achieve state-of-the-art performance. From a visual point of view, the results look convincing. Yet, does it mean that those techniques are reliable and robust enough to be implemented in real business cases to enhance the performance of other computer vision tasks? In this article, we investigate the use of SISR to construct higher-resolution images of real receipt photos sent by a company's customers and evaluate its impact on the performance of an OCR task (receipt information retrieval). Using built-in task-based performance evaluation methods, we show that the use of SISR can significantly improve OCR performance in the case where recognition was poor in low-resolution, but can also deteriorate the performance for receipts that were already successfully recognized. As a conclusion, we provide recommendations on how to best use SISR in a production environment.

Index Terms— Single image super-resolution, deep learning, task-based performance evaluation, image restoration.

1. INTRODUCTION

When dealing with images in real applications, acquisition procedures may sometime result in low-resolution (LR) images, which are an imperfect rendering of reality. Single-Image Super-Resolution (SISR) aims at constructing higher-resolution (HR) images from observed LR images which should be as similar as possible to reality. Over the last few years, AI researchers have started to use powerful deep learning (DL) algorithms for SISR. Many different methods have been used, ranging from Convolutional Neural Networks (CNN) [1, 2, 3, 4] to Generative Adversarial Networks (GAN) [5, 6, 7]. There are some issues with the current approaches to SISR. First, these algorithms mainly focus on improving the visual and perceptual image quality. From a human visual point of view, the generated images look cleaner and do feature greater resolution. Another problem is that SISR is usually trained and evaluated on synthetic datasets, where the LR images are deduced from the HR ones by down-sampling, and sometime adding various amounts of artificial noise. As pointed out recently in [8], this is not realistic, and current SISR algorithmic performances do not translate to real-world applications. The authors of [8] have proposed to build LR/HR pairs by acquiring images with different cameras and various levels of focal lengths.

Here, we evaluate the effectiveness of SISR by a different mean, specifically, can we improve the performance of other computer vision tasks by increasing the resolution of the input images with a SISR model? It is not at all obvious as the super-resolved images are not always similar to the true HR images. Indeed, they can suffer from the creation of unwanted patterns and hallucinating artifacts [9], which can lead to misinterpretation and errors. While the benefits of adopting a SISR pre-processing step have been demonstrated on four popular vision tasks for LR images, it has up to now not been shown to provide benefits on a real vision application [10]. In this article, we provide a real production use case in which SISR is applied to improve a computer vision process and evaluate the usefulness of this approach to meet a company business need.

The use case is based on a loyalty program called Scanobar and launched by the company Heineken. The main idea of this program is to reward customers when they buy Heineken products in supermarkets or bars via the Scanobar application. The latter is a chatbot available on Facebook Messenger. Customers send receipt pictures containing Heineken products to the Messenger chatbot. A computer vision algorithm detects the products and related prices on the receipts and credits a certain amount of loyalty points proportional to the price of the Heineken purchases. Google Vision's OCR is then used to extract the information (products and prices) on the received receipts. This extraction is a critical step to award the right number of loyalty points to the customers. Thus, the readability of the receipt images should be good enough so that the OCR software can read the information properly. Yet, the readability of the receipt images can be poor for 3 main reasons:

1. There is no uniformity in the pictures dataset as image sizes range from less than 5 kiB up to over 3 MiB. This can lead to information loss, especially when one needs

to resize all the images.

2. The photographs are not professionally taken or framed. They can be blurred due to motion, lack of illumination or malfunctioning autofocus, with numerous artifacts such as shadows or flash reflections. As it is typically difficult to take the whole receipt in the same picture because it is often too big to fit. As a result, the photographs is taken at a large distance from the receipts, decreasing readability significantly.

3. Finally, the photos are often heavily compressed. The pictures may be downgraded automatically when downloaded from the Scanobar bot during peak hours, resulting in much lower picture resolutions.

While there are many different causes behind the poor quality of the received pictures, in many cases, the pictures remain readable from a human point of view. However, in about 30 percent of the cases, the receipt images are readable for a human, but the OCR fails to extract the relevant information.

In this work, we develop a task-based SISR solution for the Scanobar program. It aims at improving the OCR product and price detection by enhancing the resolution of the input images which have been degraded by the 3 causes described above. We assume that the images which are not readable for humans are out of scope as they are simply too degraded. Thus, the real target is to increase the resolution of the images that are readable for humans but not for the OCR. The question is then: does the SISR model help OCR recognize information on those receipt images?

2. MODEL

Even though many state-of-the art super-resolution results on natural images have been produced using GANs, in our work we do not use such a solution due to a lack of data. Moreover, a GAN model may lack stability and can be highly sensitive to minor hyper-parameter changes as well as degradation models. From a production perspective, one needs a fast and robust model capable of generating simple patterns (figures and characters). This is why we use a simpler model based on two learning strategies, content and texture losses, also called perceptual loss in the original paper [11]. Its main objective is to enhance the images by "decrappifying" [12] them, meaning that the algorithm removes noises and increases resolution. It corresponds well to the Scanobar use case where multiple image degradation types exist. As described in [12], there are two main steps in the algorithm:

1. The generative model (gen) is a U-NET with an encoder based on a Resnet 34 backbone architecture which has been pre-trained on ImageNet. Blur is used in this model to avoid checkerboard artifacts at each layer. Cross-connection with the direct input of the model are applied (skip connection). This model takes degraded images as input and cleans them. The degraded images are constructed from HR images to which a custom degradation function D is applied.

2. For the learning strategy, one uses a VGG-16 classification model pre-trained on ImageNet. This loss network is used as inference to compare the generated images $\hat{y}$ and its corresponding target y. The idea is to be sure that the created images have the same "concept" and "style" (in our case, mainly figures and characters) than the real images. The loss function $\mathcal{L}_{model}$ is divided into 3 parts, each with its specific goal:

   - To measure the overall proximity between the 2 images, an ℓ_1 pixel loss ℓ_{pixel} is used.
   - To capture feature closeness, a content loss ℓ_{content} is expressed. We hook the activation layers of the VGG 16 intermediate blocks noted J (blocks 2 to 4) for the created image $\hat{y}$ and its corresponding target image y. Then, ℓ_1 differences are computed for each activation. We only use blocks 2 to 4 because these blocks are not specific while later blocks are too focused on the ImageNet classification task.
   - Finally, for the same feature maps J, texture loss ℓ_{texture} is used to capture style proximity. Gram Matrices are built on the feature maps J and their ℓ_1 distances are computed.

To conclude, the final loss is a combination of the ℓ_1 loss from the generative model gen and the ℓ_1 losses from the VGG activation layers J computed each time in two different ways (content and textures losses).

$$\mathcal{L}_{\text{model}} = \ell^{\text{gen}}_{\text{pixel}}(y, \hat{y}) + \ell^{\text{vgg16},J}_{\text{content}}(y, \hat{y}) + \ell^{\text{vgg16},J}_{\text{texture}}(y, \hat{y})$$

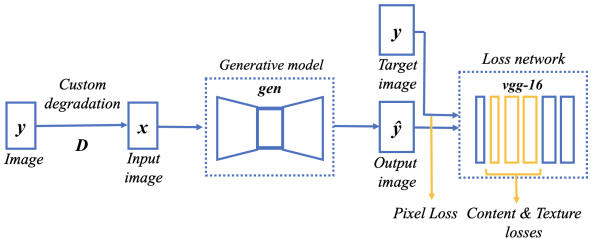


Fig. 1. System overview

3. METHOD

There are three main steps to create a useful SISR model for Scanobar: building the right training set, a representative custom degradation function and pre-processing.

3.1. Dataset

The training set should be as close as possible to the use case. Thus, *and only for training,* We used HR images of receipts which we then degraded to create the input images to our model. Fortunately, we have access to a sufficient number of real receipt images sent by customers since the launch of the application in early 2019. We combine complementary approaches:

1. We selected manually 1000 receipt images that were readable for a human and then degraded them through a custom degradation function designed to bring as close as possible to use cases (see next section). Even with data augmentation, this dataset proved too small to yield satisfying results.

2. We then added to this first dataset a set of 1000 receipt images from the ICDAR 2019 competition [13]. These receipts were much too clean to correspond exactly to our business case, but they helped with training once degraded.

3. We selected a second subset of 6000 receipt images from the Scanobar database but without any visual check on the readability of the images. In spite of this, training was still improved.

Thus, in total, the final training set is composed of 8,000 (LR, HR) pairs images to which we further apply data augmentation (random rotation, random zoom and random symmetric warp).

3.2. Degradation function

The custom degradation function applied to the images should be as similar as possible to the real degradation described in introduction: 1) we shrunk the size of the input image with bilinear interpolation, then 2) we compressed the image using the JPEG lossy standard coding; 3) finally, we applied a Gaussian Blur.

The combination of the three degradation functions corresponds to the small sensor size, default compression and poor optics of many smartphone cameras used in challenging lighting and shooting distance circumstances. In our test, a resizing to 500×500, a quality of 15 and a blur radius of 1 were the most effective.

3.3. Normalization

Normalization is important as the receipt images are heterogeneous in terms of size, brightness and contrast. In our tests, we scaled the input with a normalization based on ImageNet statistics.

4. RESULTS

4.1. Task-based evaluation method

SISR models can be evaluated using simulations, subjective testing or task-based testing.

Simulations are the most common way SISR methods are evaluated. A HR image database is degraded in some way, then enhanced via super-resolution, and the results are compared with the original HR images in terms of PSNR or SSIM.

Subjective testing is also often used. Done correctly, this can be a reliable way of evaluating image quality. Its main principle is to ask to a group of human testers their opinion on the quality of various images. Several classical methods exist to perform these studies [14, 15]. A challenge with subjective testing is to enroll a sufficient number of fair testers, as it can be time consuming, tiring and expensive.

Finally, task-based testing consists of using the SISR model to improve another computer vision task, which can in turn be evaluated objectively. The prediction of the computer vision task from original images and from SISR improved images are compared to measure the impact of SISR on the overall prediction performance. Task-based quality assessment exploit image attributes that are important for the computer vision task, not for humans. For instance, an object recognition model may focus on high-level semantics while ignoring the image contrast and noise which are two important features for humans [16]. Therefore, task-based evaluation may be useful for computer vision domain-specific applications, but less so if the end usage is to improve human perceptual quality. In our production usage case, task-based evaluation is preferable since the goal of the SISR model is to enhance OCR performance.

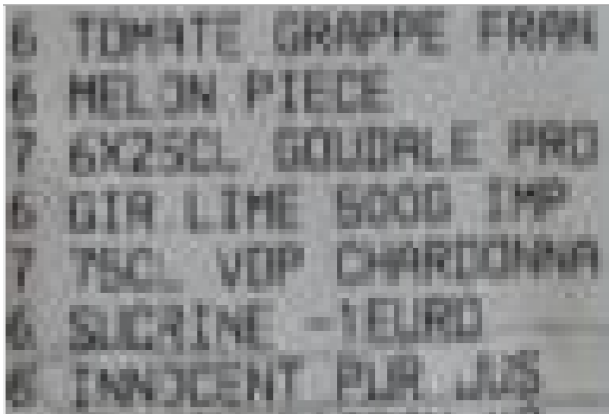
4.2. Best model results and performance evaluation

When we use our SISR model on test images (HR images which are degraded by our custom degradation function), we can see a clear visual improvement. More importantly, when applying our model to real, *non-degraded*, case images, we also observe a significant increase in quality. For example, in Fig. 2, the recognized text in the low-resolution image is entirely incorrect.

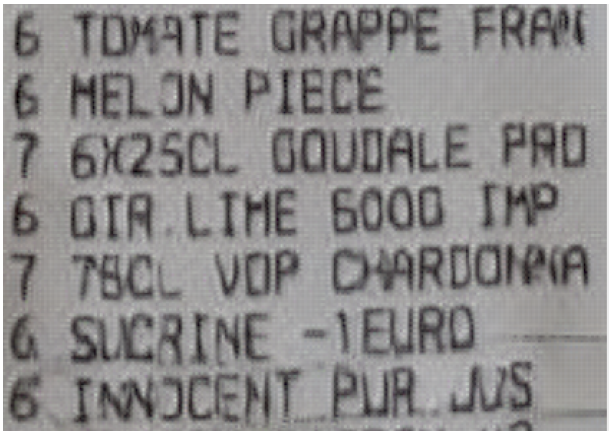
In contrast, OCR on the super-resolved text is correct in 5 out of 7 lines. It reads: "6 TOMATE GRAPPE FRAN 6 MELON PIECE ; 7 6X25CL GOUDALE PRO ; 6 OIR.LIME 6000 IMP ; 7 75CL VOP CHARDONNA ; 6 SUCRINE - 1EURD ; 6 INNOCENT PUR. JUS"

However, sometimes SISR does not help. In Fig. 3, OCR recognizes correctly the first two lines from the real LR image (top). The super-resolved image make things worse, none of the lines in the SISR image is recognized correctly from the bottom image.

To conduct the task-based evaluation of our SISR model, we need to manually label a sufficient number of receipt im-



(a)



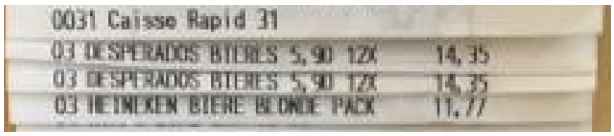
(b)

Fig. 2. (a) Zoom on a real raw image from the Scanobar database for which OCR fails to recognize the 7 item references. (b) Corresponding prediction made by the SISR model for which Google Vision OCR is far more accurate and manages to recognize perfectly 5 out of the 7 item references.

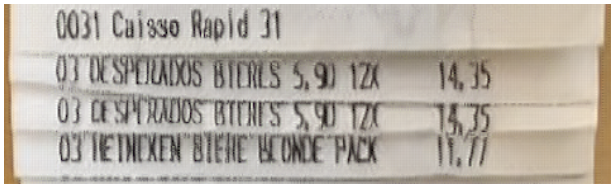
ages from the test set. We collect the following information: food store name, products, date and postal code. Performance is evaluated on 2 different datasets from the Scanobar database and is measured using the percentages of success on the total products, store names, postal codes and ticket dates additionally detected.

The first dataset (Dataset 1 in Table 1) is composed of 100 Scanobar labeled receipt images for which OCR performed poorly and failed to detect the relevant information. Results are illustrated in Table 1 and show a significant improvement for all types of information. For instance, with the SISR pre-processing step, 10% of the total Heineken products of Dataset 1 are detected additionally by the OCR. Thus, the use of the SISR on previously non-readable images (for Google Vision OCR) facilitates OCR, as illustrated on Fig. 2.

The second dataset (Dataset 2 in Table 1) contains 100 LR labeled receipts images at least partially successfully rec-



(a)



(b)

Fig. 3. (a) Zoom on a real raw image from the Scanobar database. Google Vision OCR detects perfectly the first two lines. (b) Corresponding prediction made by the SISR model. Google Vision OCR is less accurate and hardly detects any item reference.

ognized. Results in Table 1 show a generally smaller decrease in the text detected by the OCR for all types of information. Thus, we observe that SISR can sometimes worsen results.

Table 1. Impact of the SISR pre-processing step on the OCR detection performance. With SISR,

Information type	Dataset 1	Dataset 2
Heineken Products	+10%	-5%
Receipt Date	+5%	-1%
Postal Code	+15%	-9%
Food Store Name	+7%	-6%

To conclude, the SISR model is generally useful in our use case as it significantly increases the OCR performance, especially with noisy images. Even though OCR can in certain cases create pattern distortions, the gain in performance remains positive.

5. CONCLUSION

In this paper, we sought to evaluate the influence of Single-Image Super-Resolution (SISR) on OCR performance in challenging circumstances, i.e. with handheld, telephone-quality photographs of poorly printed text. We showed that text recognition increased by up to 15%. On the other hand, on some images where the recognition is already good, using SISR can degrade results by up to 9%. Thus, for production implementations, we advise to run OCR both with and without the SISR model. By keeping only the most plausible result (i.e. with correct date, real product names, etc) from the two runs, one can take advantage of the SISR ability to restore noisy receipt images while not degrading performance on clean images.

6. REFERENCES

[1] K. He C. Dong, C. C. Loy and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision*. ECCV, 2014, vol. IV, pp. 184–199.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. MICCAI, 2015, pp. 234–241.

[3] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 1646–1654.

[4] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 1874–1883.

[5] Christian Ledig, Lucas Theis, Ferenc Huszar, Ferenc Caballero, Ferenc Cunningham, Alejandro Acosta, Alejandro Aitken, Alejandro Tejani, Johannes Totz, Johannes Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv*, pp. 1–8, May 2017.

[6] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision (ECCV) Workshops*. IEEE, 2018, pp. 0–0.

[7] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," *arXiv*, pp. 1–8, June 2019.

[8] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3086–3095.

[9] S. Culley, D. Albrecht, and C. Jacobs, "Quantitative mapping and minimization of super-resolution optical imaging artifacts," *Nat Methods*, p. 263–266, February 2018.

[10] Dengxin Dai, Yujian Wang, Yuhua Chen, and Luc Van Gool, "Is image super-resolution helpful for other vision tasks?," *arXiv*, pp. 1–8, January 2016.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *arXiv*, pp. 1–8, March 2016.

[12] Jason Antic, Jeremy Howard, and Uri Manor, "Decrappification, deoldification, and super resolution," Facebook f8 conference, 2019.

[13] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar, "Icdar 2019 robust reading challenge on scanned receipts ocr and information extraction," International Conference on Document Analysis Recognition, 2019.

[14] "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT.500-11*, 2002.

[15] Pedram Mohammadi, Abbas Ebrahimi-Moghadam, and Shahram Shirani, "Subjective and objective quality assessment of image: A survey," *arXiv*, pp. 3–5, June 2014.

[16] Zhihao Wang, Jian Chen, Steven C.H. Hoi, and Fellow, "Deep learning for image super-resolution: A survey," *IEEE*, pp. 1–8, February 2019.